Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, Abhinav Gupta

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding

Gunnar A. Sigurdsson[1], Gül Varol[2], Xiaolong Wang[1], Ali Farhadi[3,4], Ivan Laptev[2], and Abhinav Gupta[1,4]

[1]Carnegie Mellon University [2]Inria [3]University of Washington [4]The Allen Institute for AI

http://allenai.org/plato/charades/

Abstract. Computer vision has a great potential to help our daily lives by searching for lost keys, watering flowers or reminding us to take a pill. To succeed with such tasks, computer vision methods need to be trained from real and diverse examples of our daily dynamic scenes. While most of such scenes are not particularly exciting, they typically do not appear on YouTube, in movies or TV broadcasts. So how do we collect sufficiently many diverse but *boring* samples representing our lives? We propose a novel Hollywood in Homes approach to collect such data. Instead of shooting videos in the lab, we ensure diversity by distributing and crowdsourcing the whole process of video creation from script writing to video recording and annotation. Following this procedure we collect a new dataset, *Charades*, with hundreds of people recording videos in their own homes, acting out casual everyday activities. The dataset is composed of 9,848 annotated videos with an average length of 30 seconds, showing activities of 267 people from three continents. Each video is annotated by multiple free-text descriptions, action labels, action intervals and classes of interacted objects. In total, Charades provides 27,847 video descriptions, 66,500 temporally localized intervals for 157 action classes and 41,104 labels for 46 object classes. Using this rich data, we evaluate and provide baseline results for several tasks including action recognition and automatic description generation. We believe that the realism, diversity, and casual nature of this dataset will present unique challenges and new opportunities for computer vision community.

1 Introduction

Large scale visual learning fueled by huge datasets has changed the computer vision landscape [1,2]. Given the source of this data, it's not surprising that most of our current success is biased towards static scenes and objects in Internet images. As we move forward into the era of AI and robotics, however, new questions arise. How do we learn about different states of objects (*e.g.*, cut vs. whole)? How do common activities affect changes of object states? In fact, it is not even yet clear if the success of the Internet pre-trained recognition models

will transfer to real-world settings where robots equipped with our computer vision models should operate.

Shifting the bias from Internet images to real scenes will most likely require collection of new large-scale datasets representing activities of our boring everyday life: getting up, getting dressed, putting groceries in fridge, cutting vegetables and so on. Such datasets will allow us to develop new representations and to learn models with the right biases. But more importantly, such datasets representing people interacting with objects and performing natural action sequences in typical environments will finally allow us to learn common sense and contextual knowledge necessary for high-level reasoning and modeling.

But how do we find these boring videos of our daily lives? If we search common activities such as "drinking from a cup", "riding a bike" on video sharing websites such as YouTube, we observe a highly-biased sample of results (see Figure 1). These results are biased towards entertainment—boring videos have no viewership and hence no reason to be uploaded on YouTube!

In this paper, we propose a novel **Hollywood in Homes** approach to collect a large-scale dataset of boring videos of daily activities. Standard approaches in the past have used videos downloaded from the Internet [3,4,5,6,7,8] gathered from movies [9,10,11] or recorded in controlled environments [12,13,14,15,16,17]. Instead, as the name suggests: we take the Hollywood filming process to the homes of hundreds of people on Amazon Mechanical Turk (AMT). AMT workers follow the three steps of filming process: (1) script generation; (2) video direction and acting based on scripts; and (3) video verification to create one of the largest and most diverse video dataset of daily activities.

There are threefold advantages of using the **Hollywood in Homes** approach for dataset collection: (a) Unlike datasets shot in controlled environments (*e.g.*, MPII [14]), crowdsourcing brings in diversity which is essential for generalization. In fact, our approach even allows the same script to be enacted by multiple people; (b) crowdsourcing the script writing enhances the coverage in terms of scenarios and reduces the bias introduced by generating scripts in labs; and (c) most importantly, unlike for web videos, this approach allows us to control the composition and the length of video scenes by proposing the vocabulary of scenes, objects and actions during script generation.

The Charades v1.0 Dataset

Charades is our large-scale dataset with a focus on common household activities collected using the Hollywood in Homes approach. The name comes from of a popular American word guessing game where one player acts out a phrase and the other players guess what phrase it is. In a similar spirit, we recruited hundreds of people from Amazon Mechanical Turk to act out a paragraph that we presented to them. The workers additionally provide action classification, localization, and video description annotations. The first publicly released version of our *Charades* dataset will contain 9,848 videos of daily activities 30.1 seconds long on average (7,985 training and 1,863 test). The dataset is collected in 15 types of indoor scenes, involves interactions with 46 object classes and has a vocabulary of 30 verbs leading to 157 action classes. It has 66,500 temporally localized actions,

Fig. 1. Comparison of actions in the Charades dataset and on YouTube: *Reading a book*, *Opening a refrigerator*, *Drinking from a cup*. YouTube returns entertaining and often atypical videos, while *Charades* contains typical everyday videos.

12.8 seconds long on average, recorded by 267 people in three continents, and over 15% of the videos have more than one person. We believe this dataset will provide a crucial stepping stone in developing action representations, learning object states, human object interactions, modeling context, object detection in videos, video captioning and many more. The dataset is publicly available at `http://allenai.org/plato/charades/`.

Contributions The contributions of our work are three-fold: (1) We introduce the Hollywood in Homes approach to data collection, (2) we collect and release the first crowdsourced large-scale dataset of boring household activities, and (3) we provide extensive baseline evaluations.

The KTH action dataset [12] paved the way for algorithms that recognized human actions. However, the dataset was limited in terms of number of categories and enacted in the same background. In order to scale up the learning and the complexity of the data, recent approaches have instead tried collecting video datasets by downloading videos from Internet. Therefore, datasets such as UCF101 [8], Sports1M [6] and others [7,4,5] appeared and presented more challenges including background clutter, and scale. However, since it is impossible to find boring daily activities on Internet, the vocabulary of actions became biased towards more sports-like actions which are easy to find and download.

There have been several efforts in order to remove the bias towards sporting actions. One such commendable effort is to use movies as the source of data [18,19]. Recent papers have also used movies to focus on the video description problem leading to several datasets such as MSVD[20], M-VAD [21], and MPII-MD [11]. Movies however are still exciting (and a source of entertainment) and do not capture the scenes, objects or actions of daily living. Other efforts have been to collect in-house datasets for capturing human-object interactions [22] or human-human interactions [23]. Some relevant big-scale efforts in

Table 1. Comparison of Charades with other video datasets.

	Actions per video	Classes	Labelled instances	Total videos	Origin	Type	Temporal localization
Charades v1.0	6.8	157	67K	10K	267 Homes	Daily Activities	Yes
ActivityNet [3]	1.4	203	39K	28K	YouTube	Human Activities	Yes
UCF101 [8]	1	101	13K	13K	YouTube	Sports	No
HMDB51 [7]	1	51	7K	7K	YouTube/Movies	Movies	No
THUMOS'15 [5]	1-2	101	21K+	24K	YouTube	Sports	Yes
Sports 1M [6]	1	487	1.1M	1.1M	YouTube	Sports	No
MPII-Cooking [14]	46	78	13K	273	30 In-house actors	Cooking	Yes
ADL [25]	22	32	436	20	20 Volunteers	Ego-centric	Yes
MPII-MD [11]	Captions	Captions	68K	94	Movies	Movies	No

this direction include MPII Cooking [14], TUM Breakfast [16], and the TACoS Multi-Level [17] datasets. These datasets focus on a narrow domain by collecting the data in-house with a fixed background, and therefore focus back on the activities themselves. This allows for careful control of the data distribution, but has limitations in terms of generalizability, and scalability. In contrast, PhotoCity [24] used the crowd to take pictures of landmarks, suggesting that the same could be done for other content at scale.

Another relevant effort in collection of data corresponding to daily activities and objects is in the domain of ego-centric cameras. For example, the Activities of Daily Living dataset [25] recorded 20 people performing unscripted, everyday activities in their homes in first person, and another extended that idea to animals [26]. These datasets provide a challenging task but fail to provide diversity which is crucial for generalizability. It should however be noted that these kinds of datasets could be crowdsourced similarly to our work.

The most related dataset is the recently released ActivityNet dataset [3]. It includes actions of daily living downloaded from YouTube. We believe the ActivityNet effort is complementary to ours since their dataset is uncontrolled, slightly biased towards non-boring actions and biased in the way the videos are professionally edited. On the other hand, our approach focuses more on action sequences (generated from scripts) involving interactions with objects. Our dataset, while diverse, is controlled in terms of vocabulary of objects and actions being used to generate scripts. In terms of the approach, Hollywood in Homes is also related to [27]. However, [27] only generates synthetic data. A comparison with other video datasets is presented in Table 1. To the best of our knowledge, our approach is the first to demonstrate that workers can be used to collect a vision dataset by filming themselves at such a large scale.

2 Hollywood in Homes

We now describe the approach and the process involved in a large-scale video collection effort via AMT. Similar to filming, we have a three-step process for generating a video. The first step is generating the script of the indoor video.

The key here is to allow workers to generate diverse scripts yet ensure that we have enough data for each category. The second step in the process is to use the script and ask workers to record a video of that sentence being acted out. In the final step, we ask the workers to verify if the recorded video corresponds to script, followed by an annotation procedure.

2.1 Generating Scripts

In this work we focus on indoor scenes, hence, we group together rooms in residential homes (*Living Room*, *Home Office*, etc.). We found 15 types of rooms to cover most of typical homes, these rooms form the scenes in the dataset. In order to generate the *scripts* (a text given to workers to act out in a video), we use a vocabulary of objects and actions to guide the process. To understand what objects and actions to include in this vocabulary, we analyzed 549 movie scripts from popular movies in the past few decades. Using both term-frequency (TF) and TF-IDF [28] we analyzed which nouns and verbs occur in those rooms in these movies. From those we curated a list of 40 objects and 30 actions to be used as seeds for script generation, where objects and actions were chosen to be generic for different scenes.

To harness the creativity of people, and understand their bias towards activities, we crowdsourced the script generation as follows. In the AMT interface, a single scene, 5 randomly selected objects, and 5 randomly selected actions were presented to workers. Workers were asked to use two objects and two actions to compose a short paragraph about activities of one or two people performing realistic and commonplace activities in their home. We found this to be a good compromise between controlling what kind of words were used and allowing the users to impose their own human bias on the generation. Some examples of generated scripts are shown in Figure 2. (see the website for more examples). The distribution of the words in the dataset is presented in Figure 3.

2.2 Generating Videos

Once we have scripts, our next step is to collect videos. To maximize the diversity of scenes, objects, clothing and behaviour of people, we ask the workers themselves to record the 30 second videos by following collected scripts.

AMT is a place where people commonly do quick tasks in the convenience of their homes or during downtime at their work. AMT has been used for annotation and editing but can we do content creation via AMT? During a pilot study we asked workers to record the videos, and until we paid up to \$3 per video, no worker picked up our task. (For comparison, to annotate a video [29]: 3 workers $\times$ 157 questions $\times$ 1 second per question $\times$ \$8/h salary = \$1.) To reduce the base cost to a more manageable \$1 per video, we have used the following strategies:
Worker Recruitment. To overcome the inconvenience threshold, worker recruitment was increased through sign-up bonuses (211% increased new worker rate) where we awarded a \$5 bonus for the first submission. This increased the total cost by 17%. In addition, "recruit a friend" bonuses (\$5 if a friend submits

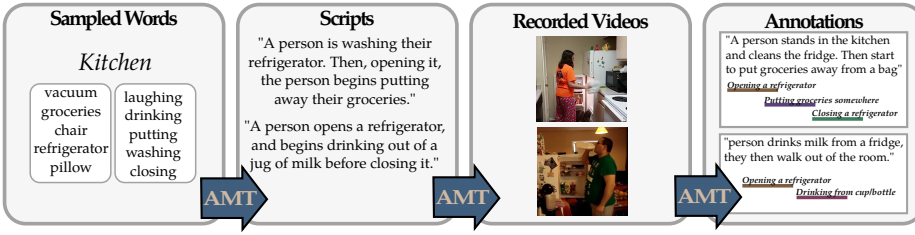

Fig. 2. An overview of the three Amazon Mechanical Turk (AMT) crowdsourcing stages in the *Hollywood in Homes* approach.

15 videos) were introduced, and were claimed by 4% of the workforce, generating indeterminate outreach to the community. US, Canada, UK, and, for a time, India were included in this study. The first three accounted for estimated 73% of the videos, and 59% of the peak collection rate.

Worker Retention. Worker retention was mitigated through performance bonuses every 15th video, and while only accounting for a 33% increase in base cost, significantly increased retention (34% increase in come-back workers), and performance (109% increase in output per worker).

Each submission in this phase was manually verified by other workers to enforce quality control, where a worker was required to select the corresponding sentence from a line-up after watching the video. The rate of collection peaked at 1225 per day from 72 workers. The final cost distribution was: 65% base cost per video, 21% performance bonuses, 11% recruitment bonuses, and 3% verification. The code and interfaces will be made publicly available along with the dataset.

2.3 Annotations

Using the generated scripts, all (verb,proposition,noun) triplets were analyzed, and the most frequent grouped into 157 action classes (*e.g.*, *pouring into cup*, *running*, *folding towel*, etc.). The distribution of those is presented in Figure 3.

For each recorded video we have asked other workers to watch the video and describe what they have observed with a sentence (this will be referred to as a *description* in contrast to the previous *script* used to generate the video). We use the original script and video descriptions to automatically generate a list of interacted objects for each video. Such lists were verified by the workers. Given the list of (verified) objects, for each video we have made a short list of 4-5 actions (out of 157) involving corresponding object interactions and asked the workers to verify the presence of these actions in the video.

In addition, to minimize the missing labels, we expanded the annotations by exhaustively annotating all actions in the video using state-of-the-art crowdsourcing practices [29], where we focused particularly on the test set.

Finally, for all the chosen action classes in each video, another set of workers was asked to label the starting and ending point of the activity in the video, resulting in a temporal interval of each action. A visualization of the data collection

process is illustrated in Figure 2. On the website we show numerous additional examples from the dataset with annotated action classes.

3 Charades v1.0 Analysis

Charades is built up by combining 40 objects and 30 actions in 15 scenes. This relatively small vocabulary, combined with open-ended writing, creates a dataset that has substantial coverage of a useful domain. Furthermore, these combinations naturally form action classes that allow for standard benchmarking. In Figure 3 the distributions of action classes, and most common nouns/verbs/scenes in the dataset are presented. The natural world generally follows a long-tailed distribution [30,31], but we can see that the distribution of words in the dataset is relatively even. In Figure 3 we also present a visualization of what scenes, objects, and actions occur together. By embedding the words based on their co-occurance with other words using T-SNE [32], we can get an idea of what words group together in the videos of the dataset, and it is clear that the dataset possesses real-world intuition. For example, *food*, and *cooking* are close to *Kitchen*, but note that except for *Kitchen*, *Home Office*, and *Bathroom*, the scene is not highly discriminative of the action, which reflects common daily activities.

Since we have control over the data acquisition process, instead of using Internet search, there are on average 6.8 relevant actions in each video. We hope that this may inspire new and interesting algorithms that try to capture this kind of context in the domain of action recognition. Some of the most common pairs of actions measured in terms of normalized pointwise mutual information (NPMI), are also presented in Figure 3. These actions occur in various orders and context, similar to our daily lives. For example, in Figure 4 we can see that among these five videos, there are multiple actions occurring, and some are in common. We further explore this in Figure 5, where for a few actions, we visualize the most probable actions to precede, and most probable actions to follow that action. As the scripts for the videos are generated by people imagining a boring realistic scenario, we find that these statistics reflect human behaviour.

4 Applications

We run several state-of-the-art algorithms on Charades to provide the community with a benchmark for recognizing human activities in realistic home environments. Furthermore, the performance and failures of tested algorithms provide insights into the dataset and its properties.

Train/test set. For evaluating algorithms we split the dataset into train and test sets by considering several constraints: (a) the same worker should not appear in both training and test; (b) the distribution of categories over the test set should be similar to the one over the training set; (c) there should be at least 6 test videos and 25 training videos in each category; (d) the test set should not be dominated by a single worker. We randomly split the workers into two groups (80% in training) such that these constraints were satisfied. The resulting

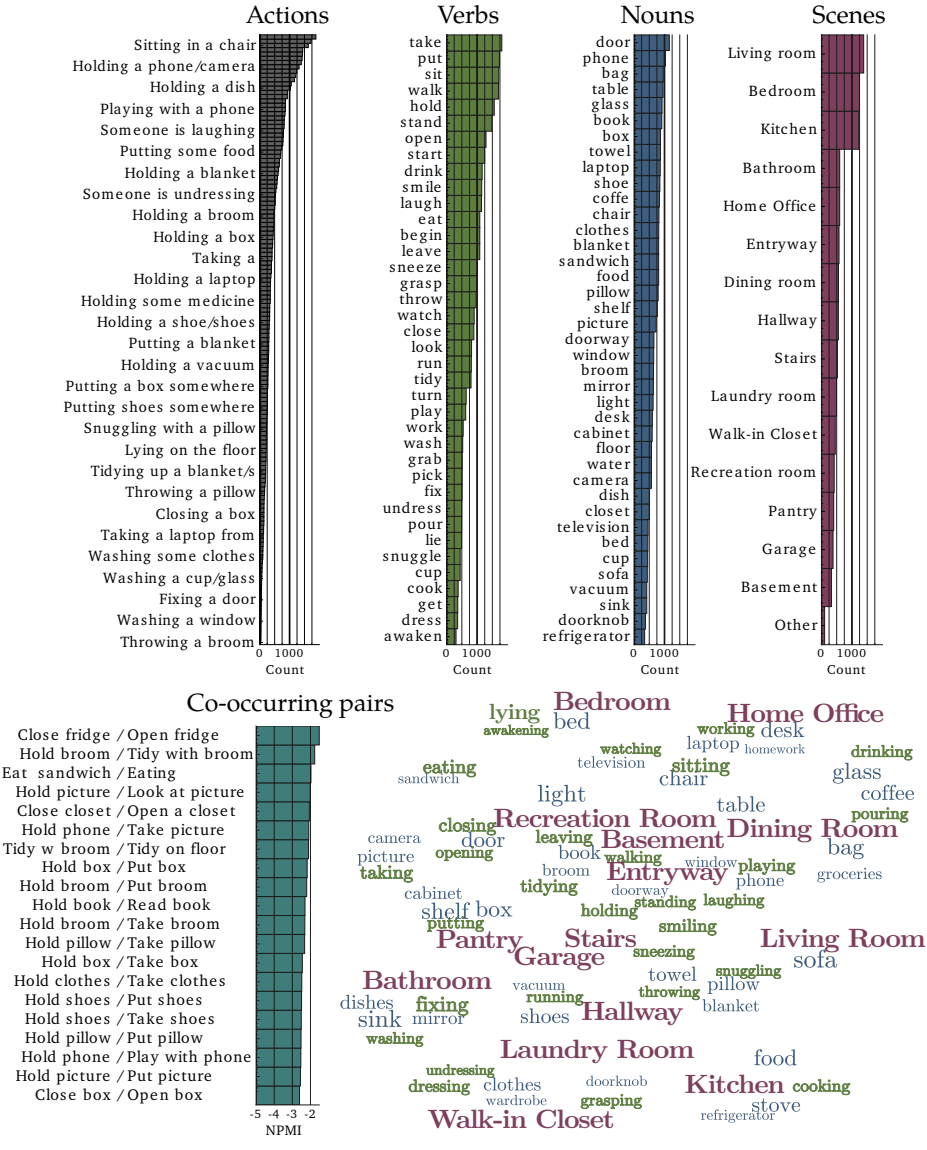

Fig. 3. Statistics for actions (gray, every fifth label shown), verbs (green), nouns (blue), scenes (red), and most co-occurring pairs of actions (cyan). Co-occurrence is measured with normalized pointwise mutual information. In addition, a T-SNE embedding of the co-occurrence matrix is presented. We can see that while there are some words that strongly associate with each other (*e.g.*, lying and bed), many of the objects and actions co-occur with many of the scenes. (Action names are abbreviated as necessary to fit space constraints.)

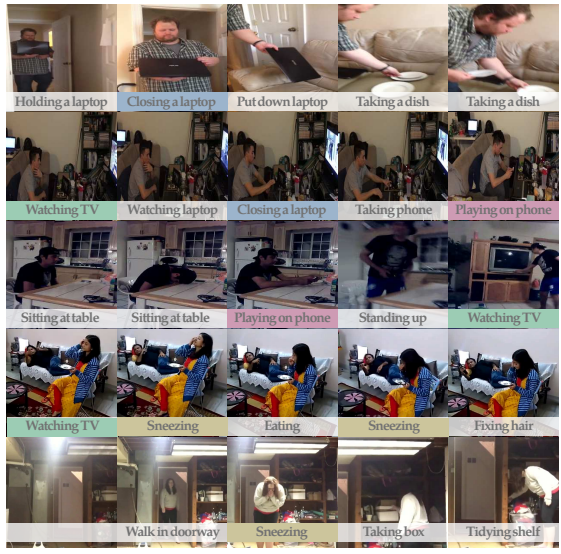

Fig. 4. Keyframes from five videos in *Charades*. We see that actions occur together in many different configurations. (Shared actions are highlighted in color).

training and test sets contain 7,985 and 1,863 videos, respectively. The number of annotated action intervals are 49,809 and 16,691 for training and test.

4.1 Action Classification

Given a video, we would like to identify whether it contains one or several actions out of our 157 action classes. We evaluate the classification performance for several baseline methods. Action classification performance is evaluated with the standard mean average precision (mAP) measure. A single video is assigned to multiple classes and the distribution of classes over the test set is not uniform.

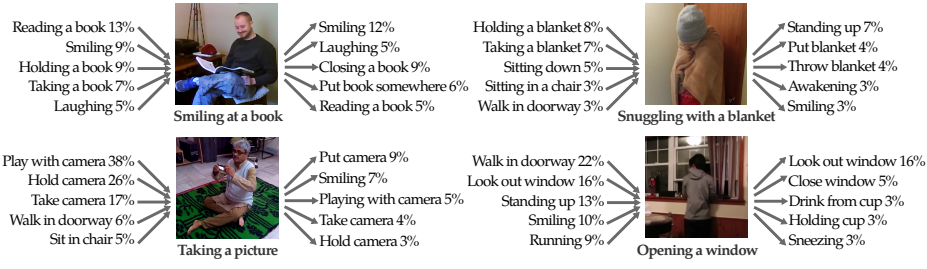

Fig. 5. Selected actions from the dataset, along with the top five most probable actions before, and after the action. For example, when *Opening a window*, it is likely that someone was *Standing up* before that, and after opening, *Looking out the window*.

Table 2. mAP (%) for action classification with various baselines.

Random	C3D	AlexNet	Two-Stream-B	Two-Stream	IDT	Combined
5.9	10.9	11.3	11.9	14.3	17.2	18.6

Table 3. Action classification evaluation with the state-of-the-art approach on Charades. We study different parameters for improved trajectories, by reporting for different local descriptor sets and different number of GMM clusters. Overall performance improves by combining all descriptors and using a larger descriptor vocabulary.

	HOG	HOF	MBH	HOG+MBH	HOG+HOF+MBH
K=64	12.3	13.9	15.0	15.8	16.5
K=128	12.7	14.3	15.4	16.2	16.9
K=256	13.0	14.4	15.5	16.5	17.2

The label precision for the data is 95.6%, measured using an additional verification step, as well as comparing against a ground truth made from 19 iterations of annotations on a subset of 50 videos. We now describe the baselines.

Improved trajectories. We compute improved dense trajectory features (IDT) [33] capturing local shape and motion information with MBH, HOG and HOF video descriptors. We reduce the dimensionality of each descriptor by half with PCA, and learn a separate feature vocabulary for each descriptor with GMMs of 256 components. Finally, we encode the distribution of local descriptors over the video with Fisher vectors [34]. A one-versus-rest linear SVM is used for classification. Training on untrimmed intervals gave the best performance.

Static CNN features. In order to utilize information about objects in the scene, we make use of deep neural networks pretrained on a large collection of object images. We experiment with VGG-16 [35] and AlexNet [36] to compute fc_6 features over 30 equidistant frames in the video. These features are averaged across frames, L2-normalized and classified with a one-versus-rest linear SVM. Training on untrimmed intervals gave the best performance.

Two-stream networks. We use the VGG-16 model architecture [37] for both networks and follow the training procedure introduced in Simonyan et al. [38], with small modifications. For the spatial network, we applied finetuning on ImageNet pre-trained networks with different dropout rates. The best performance was with 0.5 dropout rate and finetuning on all fully connected layers. The temporal network was first pre-trained on the UCF101 dataset and then similarly finetuned on conv4, conv5, and fc layers. Training on trimmed intervals gave the best performance.

Balanced two-stream networks. We adapt the previous baseline to handle class imbalance. We balanced the number of training samples through sampling, and ensured each minibatch of 256 had at least 50 unique classes (each selected uniformly at random). Training on trimmed intervals gave the best performance.

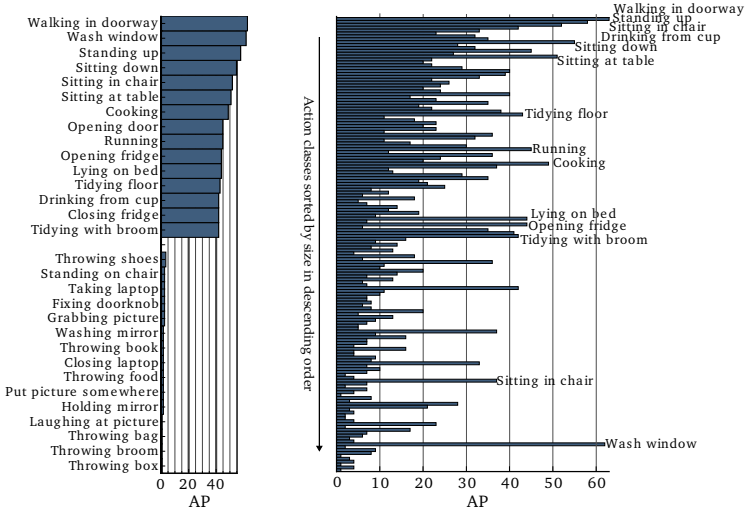

Fig. 6. On the left classification accuracy for the 15 highest and lowest actions is presented for *Combined*. On the right, the classes are sorted by their size. The top actions on the left are annotated on the right. We can see that while there is a slight trend for smaller classes to have lower accuracy, many classes do not follow that trend.

C3D features. Following the recent approach from [39], we extract fc_6 features from a 3D convnet pretrained on the Sports-1M video dataset [6]. These features capture complex hierarchies of spatio-temporal patterns given an RGB clip of 16 frames. Similar to [39], we compute features on chunks of 16 frames by sliding 8 frames, average across chunks, and use a one-versus-rest linear SVM. Training on untrimmed intervals gave the best performance.

Action classification results are presented in Table 2, where we additionally consider **Combined** which combines all the other methods with late fusion.

Notably, the accuracy of the tested state-of-the-art baselines is much lower than in most currently available benchmarks. Consistently with several other datasets, IDT features [33] outperform other methods by obtaining 17.2% mAP. To analyze these results, Figure 6(left) illustrates the results for subsets of best and worst recognized action classes. We can see that while the mAP is low, there are certain classes that have reasonable performance, for example *Washing a window* has 62.1% AP. To understand the source of difference in performance for different classes, Figure 6(right) illustrates AP for each action, sorted by the number of examples, together with names for the best performing classes. The number of actions in a class is primarily decided by the universality of the action (can it happen in any scene), and if it is common in typical households (writer bias). It is interesting to notice, that while there is a trend for actions with higher number of examples to have higher AP, it is not true in general, and actions such as *Sitting in chair*, and *Washing windows* have top-15 performance.

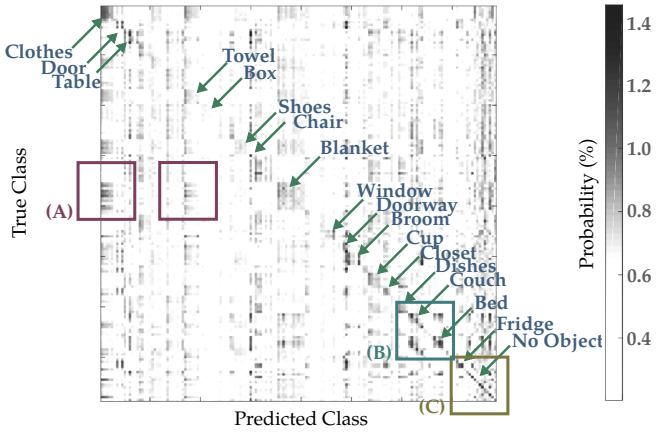

Fig. 7. Confusion matrix for the *Combined* baseline on the classification task. Actions are grouped by the object being interacted with. Most of the confusion is with other actions involving the same object (squares on the diagonal), and we highlight some prominent objects. Note: (A) High confusion between actions using *Blanket*, *Clothes*, and *Towel*; (B) High confusion between actions using *Couch* and *Bed*; (C) Little confusion among actions with no specific object of interaction (*e.g. standing up*, *sneezing*).

Delving even further, we investigate the confusion matrix for the *Combined* baseline in Figure 7, where we convert the predictor scores to probabilities and accumulate them for each class. For clearer analysis, the classes are sorted by the object being interacted with. The first aspect to notice is the squares on the diagonal, which imply that the majority of the confusion is among actions that interact with the same object (*e.g.*, *Putting on clothes*, or *Taking clothes from somewhere*), and moreover, there is confusion among objects with similar functional properties. The most prominent squares are annotated with the object being shared among those actions. The figure caption contains additional observations. While there are some categories that show no clear trend, we can observe less confusion for many actions that have no specific object of interaction. Evaluation of action recognition on this subset results in 38.9% mAP, which is significantly higher than average. Recognition of fine-grained actions involving interactions with the same object class appears particularly difficult even for the best methods available today. We hope our dataset will encourage new methods addressing activity recognition for complex person-object interactions.

4.2 Sentence Prediction

Our final, and arguably most challenging task, concerns prediction of free-from sentences describing the video. Notably, our dataset contains sentences that have been used to create the video (*scripts*), as well as multiple video *descriptions* obtained manually for recorded videos. The scripts used to create videos are biased by the vocabulary, and due to the writer's imagination, generally describe

Table 4. Sentence Prediction. In the *script* task one sentence is used as ground truth, and in the *description* task 2.4 sentences are used as ground truth on average. We find that S2VT is the strongest baseline.

	Script					***Description***				
	RW	Random	NN	S2VT	Human	RW	Random	NN	S2VT	Human
CIDEr	0.03	0.08	0.11	0.17	0.51	0.04	0.05	0.07	0.14	0.53
$BLEU_4$	0.00	0.03	0.03	0.06	0.10	0.00	0.04	0.05	0.11	0.20
$BLEU_3$	0.01	0.07	0.07	0.12	0.16	0.02	0.09	0.10	0.18	0.29
$BLEU_2$	0.09	0.15	0.15	0.21	0.27	0.09	0.20	0.21	0.30	0.43
$BLEU_1$	0.37	0.29	0.29	0.36	0.43	0.38	0.40	0.40	0.49	0.62
$ROUGE_L$	0.21	0.24	0.25	0.31	0.35	0.22	0.27	0.28	0.35	0.44
METEOR	0.10	0.11	0.12	0.13	0.20	0.11	0.13	0.14	0.16	0.24

different aspects of the video than descriptions. The description of the video by other people is generally simpler and to the point. Captions are evaluated using the CIDEr, BLEU, ROUGE, and METEOR metrics, as implemented in the COCO Caption Dataset [40]. These metrics are common for comparing machine translations to ground truth, and have varying degrees of similarity with human judgement. For comparison, human performance is presented along with the baselines where workers were similarly asked to watch the video and describe what they observed. We now describe the sentence prediction baselines in detail:

Fig. 8. Three generated captions that scored low on the CIDEr metric (red), and three that scored high (green) from the strongest baseline (S2VT). We can see that while the captions are fairly coherent, the captions lack sufficient relevance.

Random Words (RW): Random words from the training set.
Random Sentence (Random): Random sentence from the training set.
Nearest Neighbor (NN): Inspired by Devlin et al. [41] we simply use a 1-Nearest Neighbor baseline computed using AlexNet fc_7 outputs averaged over frames, and use the caption from that nearest neighbor in the training set.

S2VT: We use the S2VT method from Venugopalan et al. [42], which is a combination of a CNN, and a LSTM.

Table 4 presents the performance of multiple baselines on the caption generation task. We both evaluate on predicting the *script*, as well as predicting the *description*. As expected, we can observe that descriptions made by people after watching the video are more similar to other descriptions, rather than the scripts used to generate the video. Table 4 also provides insight into the different evaluation metrics, and it is clear that CIDEr offers the highest resolution, and most similarity with human judgement on this task. In Figure 8 few examples are presented for the highest scoring baseline (S2VT). We can see that while the language model is accurate (the sentences are coherent), the model struggles with providing relevant captions, and tends to slightly overfit to frequent patterns in the data (*e.g.*, *drinking from a glass/cup*).

5 Conclusions

We proposed a new approach for building datasets. Our Hollywood in Homes approach allows not only the labeling, but the data gathering process to be crowdsourced. In addition, *Charades* offers a novel large-scale dataset with diversity and relevance to the real world. We hope that Charades and Hollywood in Homes will have the following benefits for our community:
(1) Training data: Charades provides a large-scale set of 66,500 annotations of actions with unique realism.
(2) A benchmark: Our publicly available dataset and provided baselines enable benchmarking future algorithms.
(3) Object-action interactions: The dataset contains significant and intricate object-action relationships which we hope will inspire the development of novel computer vision techniques targeting these settings.
(4) A framework to explore novel domains: We hope that many novel datasets in new domains can be collected using the Hollywood in Homes approach.
(5) Understanding daily activities: Charades provides data from a unique human-generated angle, and has unique attributes, such as complex co-occurrences of activities. This kind of realistic bias, may provide new insights that aid robots equipped with our computer vision models operating in the real world.

6 Acknowledgements

This work was partly supported by ONR MURI N00014-16-1-2007, ONR N00014-13-1-0720, NSF IIS-1338054, ERC award ACTIVIA, Allen Distinguished Investigator Award, gifts from Google, and the Allen Institute for Artificial Intelligence. The authors would like to thank: Nick Rhinehart and the anonymous reviewers for helpful feedback on the manuscript; Ishan Misra for helping in the initial experiments; and Olga Russakovsky, Mikel Rodriguez, and Rahul Sukhantakar for invaluable suggestions and advice. Finally, the authors want to extend thanks to all the workers at Amazon Mechanical Turk.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2009) 248–255 1
2. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Neural Information Processing Systems (NIPS). (2014) 487–495 1
3. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2015) 961–970 2, 4
4. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2009) 1996–2003 2, 3
5. Gorban, A., Idrees, H., Jiang, Y.G., Roshan Zamir, A., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. `http://www.thumos.info/` (2015) 2, 3, 4
6. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2014) 1725–1732 2, 3, 4, 11
7. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: International Conference on Computer Vision (ICCV), IEEE (2011) 2556–2563 2, 3, 4
8. Soomro, K., Roshan Zamir, A., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. In: CRCV-TR-12-01. (2012) 2, 3, 4
9. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2008) 1–8 2
10. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2008) 1–8 2
11. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2015) 2, 3, 4
12. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: International Conference on Pattern Recognition (ICPR). Volume 3., IEEE (2004) 32–36 2, 3
13. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Transactions on Pattern Analysis and Machine Intelligence **29**(12) (December 2007) 2247–2253 2
14. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2012) 1194–1201 2, 4
15. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2011) 3153–3160 2
16. Kuehne, H., Arslan, A.B., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2014) 2, 4

17. Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., Schiele, B.: Coherent multi-sentence video description with variable level of detail. In: Pattern Recognition. Springer (2014) 184–195 2, 4
18. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2009) 3
19. Ferrari, V., Marín-Jiménez, M., Zisserman, A.: 2d human pose estimation in tv shows. In: Statistical and Geometrical Approaches to Visual Motion Analysis. Springer (2009) 128–147 3
20. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics (2011) 190–200 3
21. Torabi, A., Pal, C., Larochelle, H., Courville, A.: Using descriptive video services to create a large data source for video annotation research. arXiv preprint arXiv:1503.01070 (2015) 3
22. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2007) 1–8 3
23. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: International Conference on Computer Vision (ICCV), IEEE (2009) 1593–1600 3
24. Tuite, K., Snavely, N., Hsiao, D.y., Tabing, N., Popovic, Z.: Photocity: training experts at large-scale image acquisition through a competitive game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2011) 1383–1392 4
25. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2012) 2847–2854 4
26. Iwashita, Y., Takamine, A., Kurazume, R., Ryoo, M.S.: First-person animal activity recognition from egocentric videos. In: International Conference on Pattern Recognition (ICPR), Stockholm, Sweden (August 2014) 4
27. Zitnick, C., Parikh, D.: Bringing semantics into focus using visual abstraction. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2013) 3009–3016 4
28. Salton, G., Michael, J.: Mcgill. Introduction to modern information retrieval (1983) 24–51 5
29. Sigurdsson, G.A., Russakovsky, O., Farhadi, A., Laptev, I., Gupta, A.: Much ado about time: Exhaustive annotation of temporal data. arXiv preprint arXiv:1607.07429 (2016) 5, 6
30. Zipf, G.K.: The psycho-biology of language. (1935) 7
31. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annual review of neuroscience **24**(1) (2001) 1193–1216 7
32. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(2579-2605) (2008) 85 7
33. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: International Conference on Computer Vision (ICCV). (2013) 10, 11
34. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: European Conference on Computer Vision (ECCV). (2010) 10
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR). (2015) 10

36. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS). (2012) 10
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014) 10
38. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Neural Information Processing Systems (NIPS). (2014) 10
39. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: International Conference on Computer Vision (ICCV). (2015) 11
40. Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollr, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325 (2015) 13
41. Devlin, J., Gupta, S., Girshick, R., Mitchell, M., Zitnick, C.L.: Exploring nearest neighbor approaches for image captioning. arXiv preprint arXiv:1505.04467 (2015) 13
42. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: International Conference on Computer Vision (ICCV). (2015) 4534–4542 14